Minimising fourth order correlations improves latent semantic analysis performance

B. M. Pincombe[*]

(received 14 October 2005; revised 29 September 2006)

Abstract

Latent Semantic Analysis (LSA) uses semantic correlations across a corpora to reduce problems with polysemy, synonymy and inflexion when assessing document similarity. It uses singular value decomposition (SVD) to estimate a generalised linear model. This model assumes the appearance of terms in documents results from the additive noise and the product of topic and mixing matrices. Here, only the largest fourth order pairwise cross cumulants in the SVD output are minimised. Improved performance relative to LSA, as measured using precision-recall curves, is shown on the Medlars test set for a small number of retained vectors. This approach avoids the assumptions and complications of moving towards full higher order decorrelation and is also shown to produce better precision-recall curves than JADE

[*] Intelligence, Surveillance and Reconnaissance Division, Defence Science and Technology Organisation, Edinburgh, Australia.
mailto:Brandon.Pincombe@dsto.defence.gov.au

See http://anziamj.austms.org.au/V47EMAC2005/Pincombe for this article, © Austral. Mathematical Soc. 2006. Published October 26, 2006. ISSN 1446-8735

and FastICA on this standard data set. The conclusion is that minimising fourth order correlations improves the performance of LSA on at least some information retrieval tasks. Three tasks likely to benefit from removing a small number of the largest pairwise cross cumulants are identification of writing genre, detection of copied computer code, and retrieval of objects or people from video streams.

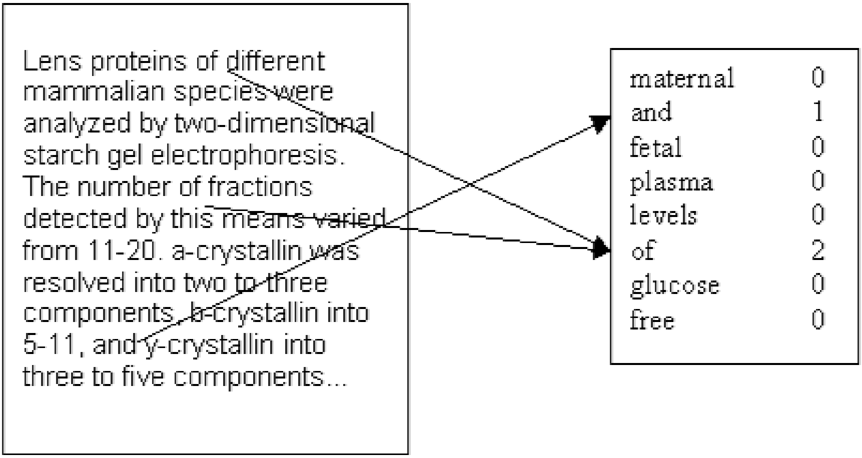
Contents

1 Introduction

Latent Semantic Analysis (LSA) provides similarity measures between text documents to allow clustering or visualisation, to identify the genre or authorship of documents, to detect plagiarism, to automatically grade student essays, to model the reading level of students in order to enable suggestion of extending texts, as a tool in assessing the psychiatric status of people with certain mental illnesses, to group voice data, and to classify video data. The immediate defence applications are likely to be limited to use in determining the similarity between documents based on their content. Done at different levels of granularity this yields information on genre, authorship and shared themes or topics of discourse. Effectively, LSA is built on the principal of removing second order correlations from the data set. Here, this principal is extended past second order to remove the largest higher order correlations.

Document sets can be represented as large term-document matrices where each document is represented as a vector of its terms. An example of this is given in Figure 1 where the number of occurrences of the words in the text characterise it as a vector. Note that the term-document matrix does not always contain a simple count of the number of terms in each document. Term-document matrices are often *weighted*. To minimise storage space and processing time they also often use a binary representation where a word is simply recorded as appearing or not appearing, but other functions of word frequency are also used. The use of term-document matrices is sometimes referred to as using a bag-of-words model as the exact position of the terms in the document is not recorded. Terms are typically words but they may be more or less complicated depending on the level of pre-processing. A more complicated example would be if an entity recogniser were run over over text containing *Australian Competition and Consumer Commission* that recognised it as a single term rather than as five words. A less complicated example is if Chinese text were represented as its constituent characters rather than as its words. In practice neither of these situations occurs often. Often uninteresting or uninformative *stop* words are left out of the set of terms used to form term-document matrices and stemming is used to attempt to deal with inflectional problems. There are numerous algorithms that estimate the similarities of documents from their term document matrices. Most of these use the first order statistics of the matrix and LSA differs from them in using second order statistics.

The common problems in determining document similarity are polysemy, synonymy and inflexion. Polysemous words have a diversity of meanings: for example, *right* can mean the side turned east when facing north, or the privilege of stockholders to subscribe to additional share issues at an advantageous price, or in accordance with what is good, proper or just, or politically conservative, or having an axis perpendicular to the base, or prompt and im-



Lens proteins of different mammalian species were analyzed by two-dimensional starch gel electrophoresis. The number of fractions detected by this means varied from 11-20. a-crystallin was resolved into two to three components, b-crystallin into 5-11, and γ-crystallin into three to five components...

maternal	0
and	1
fetal	0
plasma	0
levels	0
of	2
glucose	0
free	0

Figure 1: vector space model of text.

mediate or conformity with fact. Synonyms are two or more words sharing a meaning that is the same or nearly the same: for example, *solution* and *answer*; or *lounge*, *sofa* and *couch*. Inflexion is the process or device of adding affixes to or changing the base form of a word to express syntactic function without changing its form class. This is often done when changing tense (for example, *did*, *do* and *doing*), pluralising (for example, *dog* and *dogs*), as well as in other situations. Stemmers attempt to deal with this problem but often have difficulties with prefixes and suffixes let alone infixes (which while rare in English occur with reasonable frequency even in languages closely related to it, such as *ge* in German and Dutch, or *zu* in German) or cases where the word form changes entirely.

The use of n-grams, popular in voice recognition, has been suggested in an attempt to deal with some of these problems. n-grams are generated by sliding a window of length n through the text, for example, the 5-grams of “The quick brown fox” are “The q”, “he qu”, “e qui”, “ quic”, “quick”, “uick ”, “ick b”, ... With a felicitous choice of n some of the properties of a stemmer appear: for example, “quick”, “quickly”, “quickest” all have the 5-grams “ quic” and “quick”. For English an n of five seems best. While there are many possible n-grams, using lower case letters and space only gives 272 possible 2-grams up to 2710 possible 10-grams, in practice only a small number are used in a particular language. The vectors of n-grams can store counts or binarised data and some similarity measures can be used for n-grams as for word vectors. It is still a first order technique and still suffers many of the problems of term based first order vector space techniques with respect to polysemy, synonymy and (to a lesser extent) inflexion.

LSA is a Singular Value Decomposition (SVD) based technique for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. It uses semantic correlations across a corpora to reduce problems with polysemy, synonymy and inflexion when assessing document similarity. The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. Reduction in the problems with polysemy, synonymy and inflexion inherent in purely term based representations of documents is achieved by SVD followed by dimensionality reduction. SVD effectively estimates a generalised linear model. This model assumes that the appearance of terms in documents results from the additive noise and the product of topic and mixing matrices. The novel alteration to LSA made in this work is that (only) the largest fourth order pairwise cross cumulants in the SVD output are minimised. Improved performance relative to LSA, as measured using precision-recall curves, is shown on the Medlars test set when a small number of the singular vectors (or *factors*) found using SVD are retained to reconstruct the estimate of the term-document matrix. This approach avoids the assumptions and complications of moving towards full higher order decorrelation and is also shown to produce better precision-recall curves than JADE and FastICA on the Medlars data set (a standard test set). The conclusion is that minimising fourth order correlations improves the performance of LSA on at least some information retrieval tasks but only those in which a small number of factors are retained. Three examples of tasks likely to benefit from removing a small number of the largest pairwise cross cumulants are identification of writing genre, detection of copied computer code, and retrieval of objects or people from video streams.

2 Method

LSA uses semantic correlations across a corpora to reduce problems with polysemy, synonymy and inflexion when assessing document similarity. It starts with $\hat{\mathbf{N}}$, a $t \times d$ matrix representing d documents by their t unique index terms [1]. Local weighting, expressing the importance of a word within a document, and global weighting, scaling by the degree to which the word carries information in the domain of discourse, are applied to $\hat{\mathbf{N}}$ to produce $\mathbf{N}$ [1]. Singular Value Decomposition (SVD) is performed so that

$$\mathbf{N} = \mathbf{U}\mathbf{L}\mathbf{V}^T\,, \tag{1}$$

where the $t \times m$ matrix $\mathbf{U}$ describes the original row (or index term) entities as vectors of unit length composed of derived orthogonal factor values, the $m \times m$ matrix $\mathbf{L}$ contains the scaling (singular) values and the $m \times d$ matrix $\mathbf{V}$ describes the original column (or document) entities in the same way as $\mathbf{U}$ describes the row (index term) entries [1]. While $m = \min(t, d)$, in practice, it is unlikely that the number of documents will exceed the number of terms so $m = d$. LSA constructs $\mathbf{N}_k$, the rank k approximation of $\mathbf{N}$:

$$\mathbf{N}_k = \mathbf{U}_k\mathbf{L}_k\mathbf{V}_k^T\,, \tag{2}$$

by zeroing all but the largest k ($k \le d$) coefficients in the diagonal matrix $\mathbf{L}$. This approximation is called the latent semantic space [1].

LSA is built on the assumption that the $t \times d$ term-document matrix $\hat{\mathbf{N}}$ results from the generalised linear model

$$\hat{\mathbf{N}} = \mathbf{M}\mathbf{T} + \mathbf{E} \tag{3}$$

where $\mathbf{T}$ is a $k \times d$ topic matrix representing the distribution of the k major topics in the corpus over the d documents, $\mathbf{M}$ is a $t \times k$ mixing matrix representing the distribution of the t terms over each of the k topics, and $\mathbf{E}$ is a $t \times d$ error matrix representing noise [1, 2]. This model is less complicated than those underlying Probabilistic Latent Semantic Indexing (PLSI) [3], Latent Dirichlet Allocation (LDA) [4, 5] and the topics model [6] but still manages to closely approximate human similarity judgments [7]. LSA truncates SVD to estimate $\mathbf{T}$ in Equation (3) with $\mathbf{V}_k^T$ from Equation (2) and $\mathbf{M}$ in Equation (3) with $\mathbf{U}_k\mathbf{L}_k$ from Equation (2). This is equivalent to performing the second order decorrelation of the matrix $\hat{\mathbf{N}}$ into k principal components, with the projections $\mathbf{V}_k$ of the document vectors being uncorrelated [2], making the document vectors in $\mathbf{V}_k^T$ samples across the k independent topics represented in the corpus. We extend this principal past second order correlations to remove higher order correlations.

Assuming that the k major topics contained in $\mathbf{T}$ represent statistically independent variables, the ideal estimate of $\mathbf{T}$ should exhibit no correlations of any order amongst its columns. Were this the case Independent Components Analysis (ICA) [8] and JADE [9] would perform better than LSA. It is evident from Figure 2 that this is not the case. In place of seeking to remove all the higher order correlations between the columns of $\mathbf{T}$, the N largest fourth order inter-column correlations of $\mathbf{T}$ are minimised through orthogonal transformation of $\mathbf{U}_k$ and $\mathbf{V}_k$. Defining $\mathbf{R}$ as a $k \times k$ orthogonal matrix

$$\mathbf{R} = \prod_{(i,j)\in P_N} \mathbf{R}_{(i,j)}\,, \tag{4}$$

where $\mathbf{R}_{(i,j)}$ is a $k \times k$ orthogonal matrix rotated in only columns i and j, and P_N is the set of all pairs (i,j) such that $i \neq j$. Multiplying $\mathbf{U}_k$ and $\mathbf{V}_k$ by $\mathbf{R}$ gives the orthogonal transformations $\bar{\mathbf{U}}_k = \mathbf{U}_k\mathbf{R}$ and $\bar{\mathbf{V}}_k = \mathbf{V}_k\mathbf{R}$. The measure of the level of fourth order inter-column correlations used is the fourth order cross cumulant, defined for zero mean random variables W, X, Y and Z as

$$\mathrm{Cum}(W,X,Y,Z) \;=\; E[WXYZ] - E[WX]E[YZ]$$

$$- E[WY]E[XZ] - E[WZ]E[XY] \,.$$

Defining u_i and v_i as random variables with sample vectors comprising the columns of $\mathbf{U}_k$ and $\mathbf{V}_k$, then $\bar{v}_i = v_i \cos\phi - v_j \sin\phi$ and $\bar{v}_j = v_i \sin\phi + v_j \cos\phi$ where ϕ is the rotation angle parameterising $\mathbf{R}_{(i,j)}$. This is because multiplication of $\mathbf{U}_k$ and $\mathbf{V}_k$ by $\mathbf{R}_{(i,j)}$ changes only columns i and j.

Minimising the N largest absolute pairwise cross cumulants removes the N largest outliers from the histogram of pairwise cross cumulants leading to $\bar{\mathbf{U}}_k$ and $\bar{\mathbf{V}}_k$ being closer to sample vectors from k independent variables than are $\mathbf{U}_k$ and $\mathbf{V}_k$. This should improve performance to the extent that the independence assumption is correct. To achieve this $\mathbf{R} = \mathbf{I}_k$, $\bar{\mathbf{U}}_k = \mathbf{U}_k$ and $\bar{\mathbf{V}}_k = \mathbf{V}_k$ are used as initial conditions and then (i,j) are chosen as $\operatorname{argmax}_{(i',j')}|\operatorname{Cum}(\bar{v}_{i'}, \bar{v}_{i'}, \bar{v}_{j'}, \bar{v}_{j'})|$ and ϕ^*, the optimal angle, is selected as $\operatorname{argmin}_{\phi}|\operatorname{Cum}(\bar{v}_{i'}, \bar{v}_{i'}, \bar{v}_{j'}, \bar{v}_{j'})(\phi)|$. Then $\bar{\mathbf{U}}_k = \bar{\mathbf{U}}_k \mathbf{R}_{(i',j')}(\phi^*)$, $\bar{\mathbf{V}}_k = \bar{\mathbf{V}}_k \mathbf{R}_{(i',j')}(\phi^*)$ and $\mathbf{R} = \mathbf{R}\mathbf{R}_{(i',j')}(\phi^*)$ are calculated and the process repeated N times.

3 Results and discussion

Medlars is a standardised test collection containing 1033 medical abstracts. An example of one of these abstracts, abstract 13, reads:

> analysis of mammalian lens proteins by electrophoresis:
>
> lens proteins of different mammalian species were analyzed by two-dimensional starch gel electrophoresis. the number of fractions detected by this means varied from 11-20. a-crystallin was resolved into two to three components, b-crystallin into 5-11, and y-crystallin into three to five components. this technique provides a sensitive method for the fractionation of lens proteins and for analyzing species differences.

and is fairly typical of the contents of Medlars. The test collection also contains 30 queries. For example, query number one is

> the crystalline lens in vertebrates, including humans.

Human judges determined the abstracts that should be retrieved for each query; for example, for query 'one' there are 37 relevant abstracts including abstract 13 detailed above. Some of the abstracts in the corpus are relevant to more than one query whereas others are not relevant to any. In the results shown in this paper all 30 queries were used.

On the Medlars test collection precision-recall curves (see Figure 2) were generated for a rank twenty ($k = 20$) approximation of the term-document matrix by LSA with twenty ($N = 20$), ten ($N = 10$), five ($N = 5$) and none of the largest higher order inter-column correlations removed as well as by ICA and JADE. Note that the situation with none of the largest higher order inter-column correlations removed is simply normal LSA. The curves displayed in Figure 2 were generated by examining all 30 queries on all 1033 abstracts. Linkages between queries and abstracts were ranked according to their probability and the precision-recall curve was generated from this ordered list of probabilities.

The most evident feature of the results displayed in Figure 2 is that the curve for LSA when the largest five, fourth order, pairwise cross cumulants in the SVD output are minimised is closest to the top right corner and therefore this method has the best overall performance. Minimisation of the ten largest fourth order inter-column correlations produces results that are slightly better than ordinary LSA and minimisation of the twenty largest fourth order pairwise cross cumulants yields results that are slightly worse. JADE performs better than FastICA but neither do as well as the LSA variants.

These results seem consistent with a situation where minimisation of the largest fourth order pairwise cross cumulants improves performance but as

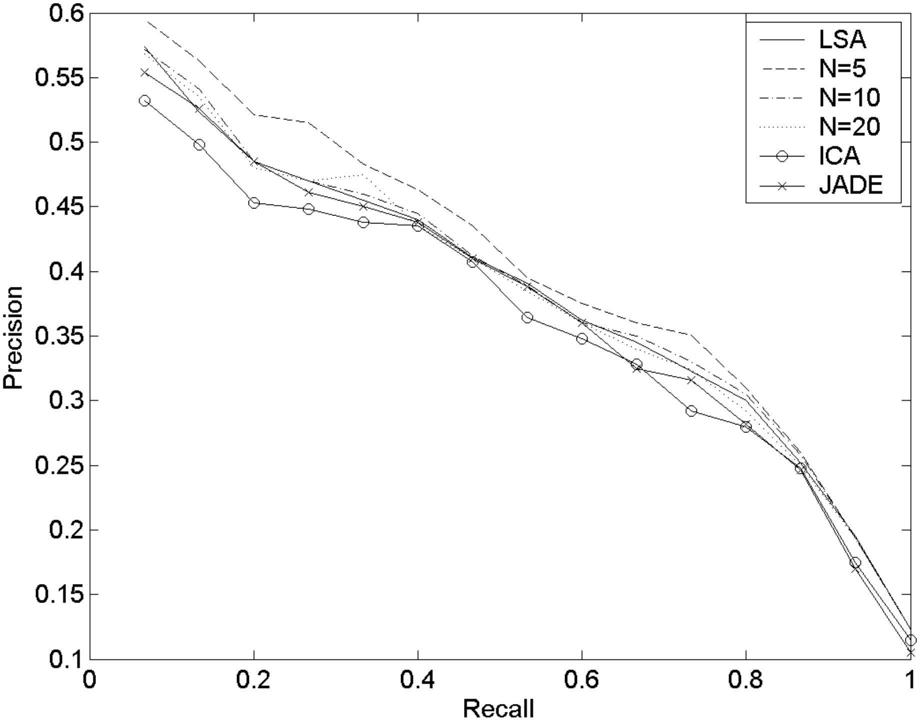


Figure 2: Precision-recall curves for LSA with 20, 10, 5 and none of the largest higher order inter-column correlations minimised as well as for ICA and JADE. In all cases $k = 20$.

smaller and smaller fourth order pairwise cross cumulants are minimised performance degrades. This means that the k major topics contained in the topic matrix $\mathbf{T}$ do not represent statistically independent variables but that there is some level of independence.

The results shown in Figure 2 are for a rank twenty ($k = 20$) approximation of the term-document matrix. As the rank of the approximation increases the ability of all techniques to accurately find topical abstracts improves and the precision-recall curve moves towards the top right corner of the graph. The various techniques also converge in their performance as k increases. This occurs first for the variations of LSA at about $k = 80$ and then for ICA and JADE at about $k = 150$.

4 Conclusion and further work

Minimisation of the five largest, fourth order, inter-column correlations has been shown to boost the performance of LSA on the Medlars standardised test set when only 20 vectors are retained. This boost in performance degrades as more vectors are retained. The removal of larger numbers of fourth order inter-column correlations yields results that are similar to simply performing LSA. Therefore only a small number of the largest higher order inter-column correlations should be removed for optimum performance and this technique should only be applied to situations where in the order of 20 dimensions are to be used. Given the requirement that the rank of the approximated term-document matrix be around twenty, situations where minimisation of the five largest higher order inter-column correlations would be appropriate include the separation of genres (for example, German prose and poetry [10]), the detection of cheating in computer assignments [11] or video object retrieval [12]. We suggest that the method be tested on problem sets from these areas.

Acknowledgements: Geoff Latham agreed to help with this work very shortly before his untimely death. He was an excellent mathematician and, had he lived a little longer, this work would undoubtedly have been completed many years ago and have a much more rigorous phrasing. Thanks are also due to John Asenstorfer, Marcus Butavicius, Leon Casey, Simon Dennis, Michael Lee, Carey Priebe, Julian Sorensen and Chris Woodruff for moral or practical support.

References

[1] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Fernas and R. A. Harshman, Indexing by latent semantic analysis, *Journal of the American Scoiety of Information Science*, vol. 41, pp. 391–407, 1990. C424, C425

[2] C. H. Papadimitrou, P. Raghavan, H. Tamaki and S. Vempala, Latent semantic indexing: a probabilistic analysis. In *Proc. ACM Conference on Principles of Database Systems*, Seattle, WA, 1998, pp. 159–168. C425

[3] T. Hofmann, Probabilistic latent semantic indexing. In *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, USA, 1999, pp. 50–57. C425

[4] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation. In *Neural Information Processing Systems 14*, MIT Press, Cambridge, MA, 2002, pp. 601–608. http://portal.acm.org/citation.cfm?coll=GUIDE&dl=GUIDE&id=944937 C425

[5] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. C425

[6] T. L. Griffiths and M. Steyvers, A probabilistic approach to semantic representation. In *Proc. of the Twenty-Fourth Annual Conference of Cognitive Science Society*, George Mason University, Fairfax, VA, 2002. C425

[7] M. D. Lee, B. Pincombe and M. Welsh, An empirical evaluation of models of text document similarity. In *Proc. of the XXVII Annual Conference of the Cognitive Science Society*, pp. 1254–1259, 2005. URL C425

[8] A. Hyvärinen, J. Karhunen and E. Oja. *Independent component analysis*. John Wiley & Sons, New York, 2001. C425

[9] J. F. Cardoso and A. Souloumiac, Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, December 1993. C425

[10] P. Nakov, Latent semantic analysis for German literature investigation. In *Proceedings of the 7th Fuzzy Days'01, International Conference on Computational Intelligence*, LNCS 2206, pp. 834–841, 2001. C429

[11] P. Nakov, Latent semantic analysis of textual data. In *Proceedings of the International Conference on Computer Systems and Technologies*, pp. V.3-1–V.3-5, 2000.
`http://portal.acm.org/citation.cfm?id=365382` C429

[12] F. Souvannavong, L. Hohl, B. Merialdo and B. Huet, Structurally enhanced latent semantic analysis for video object retrieval. *IEE Proceedings Vision, Image and Signal Processing*, vol. 152, no. 6, pp. 859–867, 2005. C429